Evaluation	Grades Evaluated	Effect Sizes	
Abt evaluation found positive effects in reading and math in most sites; JDRP-approved sites include Flint, Mich.; Dayton, O., Flippin, Ark.; E. Las Vegas, N.M.; Uvalde, Tex.; Kingstree, S.C.; Cherokee, N.C.; and Washington, D.C.	Grades K-3 (4 yrs.)	Reading comp. Language Computation Problem solving	+.07 +.84 +.57 +.17
Both time-series and control-group comparisons showed clear effects in reading and math in suburban school near Salt Lake City. *Not disadvantaged.*	Grades 3-5 (2 yrs.)	Reading Math	+.45 +.27
Matched control-group design found convincing effects in reading in rural Princeton, Ill.; some students were Title I, but most were *not disadvantaged.*	Grade 2 +.72 Grade 3 +.80 Grade 4 +.50 Grade 5 +.39 Grade 6 +.43	Reading vocab. Reading comp.	+.53 +.61
Most evaluations were fall/spring or spring/spring. One control-group evaluation in suburban Chapter 1 school near Cincinnati found convincing effects.	Grade 4	Reading vocab. Reading comp.	+.51 +.39
Study compared students in Lincoln, Neb., schools that had successfully implemented program to matched schools. *Not disadvantaged.*	Grade 2	Word knowledge Reading	+.30 +.23
Study compared project schools with matched controls in suburb of Salt Lake City. *Not disadvantaged.*	Grade 1 +.11 Grade 2 .00 Grade 3 +.41 Grade 4 +.23 Grade 5 +.44 Grade 6 +.39	Reading vocab. Reading comp.	+.39 +.14
Study compared project schools to matched controls in Miami, Fla.	Grade 1	Paragraph meaning Wide Range Achievement (WRAT)	+.95 +.28

their Chapter 1 classmates. (The one exception is the small proportion of schools at least 75% of whose students are in poverty; under certain circumstances, such schools may implement schoolwide projects in which all students may benefit from Chapter 1 funds.) The restriction of Chapter 1 funds to only those students eligible to receive them creates an interesting paradox. Legislation intended to concentrate resources on low-achieving, disadvantaged students may inhibit the use of those classroom programs that have the greatest potential for meeting their needs, as well as for helping higher-achieving disadvantaged students reach their full potential.

INTENSIVE SUPPLEMENTARY PROGRAMS

Intensive supplementary models are a second major category of effective programs. These programs offer high-quality instruction for relatively short periods, usually in addition to regular classroom instruction. There are three major types of intensive supplementary programs: preventive tutoring models, remedial tutoring models, and computer-assisted instruction. Information on effective programs of these types is summarized in Tables 3, 4, and 5.

Preventive tutoring programs. Probably the most effective (but also the most expensive) of the various programs identified in our review were models in which specially trained adult tutors worked one-on-one with at-risk first-graders in an attempt to insure their success in initial reading (see Table 3). The idea behind these programs is to prevent the development of early reading problems, thereby making later remedial and special education unnecessary. Such programs typically produce substantial gains in reading by the end of first grade, and in a few studies these gains have been sustained at least through the end of second grade.[15]

Remedial tutoring programs. Programs that use adult volunteers or older students to tutor low-achievers are a second category of tutoring models. Because they rely on unpaid tutors, the costs of such programs are low, yet their effects can be substantial (see Table 4). In addition, tutoring programs that use older students as tutors tend to produce achievement gains for the tutors as well as for the tutees.[16]

Computer-assisted instruction. Several computer-assisted instruction (CAI) programs have been found to be instructionally effective (see Table 5). The most widely evaluated and consistently successful of these have been forms of

Neither federal nor state governments have any business mandating schools to use particular programs.

TABLE 2.
Cooperative Learning Programs

Program and Source	Grades/ Subjects	Description
Team Accelerated Instruction (JDRP #84-5)	3-6 Math	Students worked on programmed materials in mixed-ability teams while teachers taught same-ability teaching groups.
Cooperative Integrated Reading and Composition	3-5 Reading Writing	Students worked in mixed-ability teams while teachers taught reading groups.

Source: Robert E. Slavin and Nancy A. Madden, "Effective Classroom Programs for Students at Risk," paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., 1987.

the Computer Curriculum Corporation's drill-and-practice programs, on which students spend about 10 minutes per day in addition to their usual reading and math instruction.[17] Successful CAI models tend to be very expensive, and there is some question about whether the computer itself is a necessary ingredient. For example, the best evaluation of Basic Literacy Through Microcomputers (JDRP No. 84-14) used typewriters rather than computers to supplement the reading instruction of first-graders.

Unlike the continuous-progress and cooperative learning programs, tutoring or CAI programs may be funded under existing Chapter 1 rules, because these are pull-out programs that can be restricted to Chapter 1 students.

POLICY IMPLICATIONS

The programs discussed in the previous section certainly do not represent the total universe of instructional programs that have been found to be effective for elementary students, and the programs that appear in Tables 1 through 5 may in the future prove to be less effective than they now seem to be. However, let us assume for a moment that it is possible, now or in the future, to assemble a list of instructional programs that are widely acknowledged to be markedly more effective than traditional pull-out or in-class Chapter 1 models. The consequences could be revolutionary.

Chapter 1 legislation currently requires school districts applying for funds to guarantee their state education agencies that their programs "are of a sufficient size, scope, and quality to give reasonable promise of substantial progress toward meeting the educational needs of the children being served."[18] Even though research on existing Chapter 1 pull-out programs has found some achievement benefits for at-risk students, these are hardly large enough to be called "substantial progress." If we have available instructional programs that *do* give reasonable promise of substantial progress, then the political and legal calculus surrounding Chapter 1 programs is significantly altered.

Although legislation authorizing Title I and Chapter 1 has never mandated the use of pull-outs, school districts have overwhelmingly used pull-out models, largely because they provide the safest means of insuring that only students eligible for Chapter 1 are served by Chapter 1 funds.[19] If it can be shown that alternatives to traditional models are considerably more effective, then state education agencies may have the right, under the principle of "substantial progress," to demand that local districts stop submitting Chapter 1 proposals based on ineffective pull-out models or in-class models that amount to the same thing. Neither federal nor state governments have any business mandating particular programs. But they could demand that school districts choose from among coherent programs with proven track records for improving the achievement of students at risk of school failure.

In essence, if we know that we *can* make a substantial difference in the achievement of low-performing students, then we have a moral imperative to see that we *do* make that difference. But how can we bring about a state of affairs in which school districts are choosing from among proven programs, implementing these programs effectively, and producing measurable benefits for their at-risk

"Mr. Fristo told me today that the school's not big enough for both of us. Isn't that a funny thing to say about a huge place like Wilson Elementary?"

Evaluation	Grades Evaluated	Effect Sizes	
In three studies — two with random assignment, one matched — TAI classes exceeded controls. One study was in rural Md., one in suburban Md., and one in urban Del.	Full sample (grades 3-6)	Computation	+.50
		Math concepts and applications	+.06
	Mainstreamed Students	Computation	+.38
		Math concepts and applications	+.47
CIRC classes exceeded matched control classes in suburban Baltimore study.	Full sample (grades 3-4)	Reading vocab.	+.12
		Reading comp.	+.35
		Oral reading	+.54
		Language	+.30
	Remedial Students	Reading vocab.	+.26
		Reading comp.	+.40
		Oral reading	+.71
		Language	+.37
	Mainstreamed Students	Reading vocab.	+.90
		Reading comp.	+.99
		Language	.00

students? Doing this on a national scale would require many changes in legislation, funding, and practice. These changes are outlined in the following sections.

1. *Establish a valid list of effective programs for at-risk students.* The first thing we need to do is to compile a set of existing instructional programs that show substantial evidence of instructional effectiveness. Distinguishing programs with rigorously evaluated, consistent effects from those with less-convincing evidence of success is essential. It will enable school districts to be reasonably sure that their investment in a particular program will pay off in gains in student achievement.

A mechanism that already exists for identification of effective programs is the Joint Dissemination Review Panel, which I mentioned above. The JDRP reviews programs developed under federal funding to see if they present adequate evidence of effectiveness. Programs that pass JDRP review are eligible for funding and dissemination through the National Diffusion Network.

The JDRP/NDN process is a conceptually sound one, but it has many flaws in practice. First, the panel has tended to look for programs with large effects and has paid less attention to experimental design. This has led to the acceptance by the JDRP of many projects that used as criteria for program effectiveness fall-to-spring percentile gains or normal-curve-equivalent gains, both of which greatly exaggerate program effectiveness. The majority of the more than 400 JDRP-approved programs used such designs.[20] Second, the evaluations presented to the JDRP may represent atypically effective sites rather than generally effective programs. Third, many programs approved by the JDRP were developed to meet local needs and do not lend themselves well to replication elsewhere.

What we need is a much shorter list of proven instructional programs that could be used effectively in Chapter 1 programs. Tables 1 through 5 represent a first step in this direction – but much more remains to be done.

2. *Provide resources to help districts adopt effective practices.* Once we have a list of effective models that could be used to meet the needs of students at risk, we will need a mechanism for implementing these programs successfully in Chapter 1 schools. There are many ways to arrange this.

For example, state or regional Chapter 1 Effectiveness Centers might be established, staffed by personnel trained in the various effective models and in the dissemination and implementation of successful practices. These centers could conduct intensive staff development programs with teams of teachers and administrators from various districts, as in California's highly successful Multi-District Teacher Training Institutes.[21] As an alternative to maintaining a fully staffed center, state-level agencies might be established to bring developers of effective programs into local districts, much as the NDN's state facilitators already do. At the local level, some proportion of Chapter 1 funds (or perhaps Chapter 2 funds) could be set aside to pay for training, follow-up, and evaluation of effective models.

3. *Fund research and development of new effective models.* Most of the effective programs listed in Tables 1 through 5 are quite old, having been developed in the 1960s and early 1970s when federal money for program development was available. However, since the Nixon Administration there has been little federal support for the kinds of comprehensive development and evaluation of instructional programs that have the greatest promise for making substantial differences in the achievement of at-risk students. It is important to provide school districts with a wide range of effective programs from which to choose and to make a substantial commitment to a serious, long-term effort to translate the knowledge about teaching and learning accumulated over the past 20 years into coherent, comprehensive instructional programs for students at risk. Billions have been spent on Chapter 1 activities,

Needed: a sizable, long-term commitment to the creation of effective programs for at-risk students.

millions on Chapter 1 evaluation, but not one cent has been spent on development and evaluation of research-based programs specifically designed to use Chapter 1 resources to prevent or remediate learning problems.

What is needed is a sizable, long-term commitment to the creation of effective programs for at-risk students. Several groups of researchers and developers should be funded to follow a rational sequence of development, pilot-testing, small-scale evaluation, large-scale evaluation by developers, and large-scale evaluation by independent evaluators. Some projects already in existence might be ready to begin at some point in the middle of this sequence, while others might need to begin at the conceptualization stage. The idea, though, would be to put the very best minds in educational research and development (and related fields) to work in pursuit of a well-defined objective of enormous importance: the prevention and remediation of learning problems among students at risk of school failure.

In addition to a need for substantial and long-term funding, another essential element of such an effort is that the schools participating in research on effective programs for students at risk must be exempted from many Chapter 1 restrictions, particularly the requirement that Chapter 1 funds be spent only on students designated as eligible for the program. In fact, it might be a good idea to establish a set of criteria for research design (e.g., use of control groups, adequate measures) and allow any school district that is interested in evaluating innovative programs and is willing to meet these criteria to receive a waiver of certain Chapter 1 restrictions.

For example, Chapter 1 schools that do not meet the current 75% poverty requirement for implementation of schoolwide programs might be allowed to implement innovative schoolwide projects anyway if the research design allowed the schools to assess the achievement effects of the innovative schoolwide project for students eligible for Chapter 1. With certain safeguards, such a system could encourage state and local education agencies to set up well-designed experiments to test plausible programs that could supplement federal efforts. Furthermore,

TABLE 3.
Preventive Tutoring Programs

Program and Source	Grades/ Subjects	Description
Programmed Tutorial Reading (JDRP #74-17)	1 only Reading	Each student was tutored 15 minutes per day by paraprofessionals or older students.
Prevention of Learning Disabilities — New York (JDRP #79-33)	1-2 Reading	Each student was tutored three to five times per week for 30 minutes by a resource teacher.
Wallach Tutorial Program	1 only Reading	Students tutored by paraprofessionals for 30 minutes per day, 28 hours total.
Reading Recovery	1 only Reading	Students were tutored by specially trained teachers for 30 minutes per day, at least 60 days.

Source: Nancy A. Madden and Robert E. Slavin, "Effective Pull-Out Programs for Students at Risk," paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., 1987.
*Grade equivalents.

TABLE 4.
Remedial Tutoring Programs

Program and Source	Grades/ Subjects	Description
Training for Turnabout Volunteers (JDRP #81-11)	Tutors 7-9 Tutees 1-6 Reading Math	This was a cross-age tutoring program. Tutors were trained with preservice and inservice classes. The material was not programmed. Students were tutored 40 minutes a day, four days a week. The fifth day was used for inservice training.
School Volunteer Development Project (JDRP #75-79)	2-6 Reading Math	Community volunteers tutored students for two to four hours per week; each student was tutored for a minimum of one-half hour, four days a week.
Success Controlled Optimal Reading Experience (SCORE) (JDRP #80-42)	1-6 Reading Math	Each student was tutored for 15 minutes per day by older students or adult volunteers.

Source: Nancy A. Madden and Robert E. Slavin, "Effective Pull-Out Programs for Students at Risk," paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., 1987.

Evaluation	Grades Evaluated	Effect Sizes		
Evaluation contained matched pairs of tutored/nontutored students in Farmington, Ut. There were no pretest differences. Effects were replicated using non-standardized tests in several other districts.	Grade 1 (bottom quartile)	Vocab. Comprehension Basal comprehension measure		+.65 +.41 +.50
Students were randomly assigned to tutoring or control treatments. Achievement was measured at end of first and second grades.	Grades 1 and 2 (students with low readiness scores)		Gr. 1	Gr.2
		WRAT	+.86	+1.06
		Word attack	+1.41	+1.67
		Word recog.	+.95	+.91
Unclear how students were assigned to treatments, but groups were equivalent at pretest. Compared tutoring to no tutoring.	Grade 1 (students below 40% on Metropolitan Achievement Tests [MAT])	Spache Word Recognition California Test of Basic Skills (CTBS) (total reading)		+1.3 * +.75
First study compared lowest readers in 14 Columbus, O., classes to similar students in comparison classes. Second study randomly assigned low readers to experimental and control groups.	Grade 1 (low readers)	CTBS (reading vocab.) CTBS (reading comp.)		+.70 +.92

Evaluation	Grades Evaluated	Effect Sizes	
One-year study in Miami, Fla., compared trained vs. untrained tutors in nonequivalent but unbiased groups. Five schools took part in the study.	Tutees 1-6	Tutees: MAT (math) MAT (reading)	 +.93 +.51
	Tutors 7-9	Tutors: MAT (math) MAT (reading)	 +.49 +.08
Students in Miami, Fla., were randomly assigned to tutored or nontutored conditions for one school year.	Grades 2-6	MAT (reading) MAT (math)	+.50 +1.1
Students in San Francisco were randomly assigned to tutored or control conditions — two controls were replicated over three years. No tests of reading comprehension were conducted; effects are for word recognition only.	Grades 1-6	WRAT Gilmore	+.5 -.7 +.5

such a change would allow school districts to use Chapter 1 funds for classroom programs that benefit all students only if they could show that students within these classes who are eligible for Chapter 1 are making substantial progress.

4. *Establish independent evaluation centers*. Evaluations of programs by external, independent evaluators would be essential to insure the credibility of program effects. If we are to invest heavily in a small number of programs, we must be able to place a great deal of faith in their evaluations. The federal government should fund one or more independent evaluation centers, whose job would be overseeing and assisting in independent program evaluations conducted by state and local education agencies, as well as conducting evaluations of their own. In these evaluations, the outcome measures used would not be made available to the developers in advance (although

A penny on the dollar would mean $39 million allocated solely to research and development.

developers would be told what domains the evaluations would assess). Teachers or schools might be randomly assigned to project or control conditions across a variety of sites.

These independent evaluations are a critical final step in the sequence of development and evaluation described above. Among other things, the knowledge that their products would ultimately be subjected to independent evaluation on valid measures of achievement that are not specifically keyed to the developer's own approach would force developers to conduct high-quality evaluations themselves and to avoid designing instructional strategies that essentially "teach to the test."

Developers would submit their programs to the independent evaluation centers to be considered for evaluation in a given year. The centers (or other review panels) would decide which programs to evaluate first by considering both the evaluation data submitted by the developer and the plausibility of and need for the program.

Obviously, it would be several years before newly developed programs could be ready for independent evaluation. However, the independent evaluation centers could begin by evaluating the most promising of the existing programs, such as those listed in Tables 1 through 5.

TABLE 5.
Computer-Assisted Instruction Programs

Program and Source	Grades/ Subjects	Description
Computer Curriculum Corp. Study 1, LAUSD Study	1-6 Math 3-6 Reading, Language	Students engaged in drill-and-practice activities on computers for 10 minutes per day in addition to regular instruction.
Computer Curriculum Corp. Study 2, Lafayette Parish Title I Math (JDRP #82-46)	3-6 Math (Chapter 1)	Students engaged in drill-and-practice activities on computers for 10 minutes per day in addition to regular instruction.
Computer Curriculum Corp. Study 3, Merrimack Education Center (JDRP #82-34)	2-9 Reading (Chapter 1)	Students engaged in drill-and-practice activities on computers for 10 minutes per day in addition to regular instruction.
Basic Literacy Through Microcomputers (JDRP #84-14)	1-3 Reading (Chapter 1)	Microcomputers or typewriters were used to practice phonics and writing skills in addition to regular instruction.

Source: Nancy A. Madden and Robert E. Slavin, "Effective Pull-Out Programs for Students at Risk," paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., 1987.

A PENNY PROPOSAL

Imagine that we had a set of programs that we knew to be instructionally effective (if properly implemented) for students at risk of school failure. Imagine that we knew what was required to implement and maintain these programs and that adequate resources were available to provide the training, support, and follow-up needed to implement them. If such a happy state of affairs existed, the situation with respect to the education of students who are at risk of school failure would be altered dramatically. We would no longer have excuses for failing to provide the most effective interventions for the neediest students; to do so would be as unacceptable as withholding effective medical treatment from children with curable diseases.

To bring this situation about, all that is needed is a long-term commitment of funds to get the best minds working on the problem. It is shameful that in the $3.9 billion Chapter 1 program there is *nothing* allocated to insuring that the programs bought at such cost actually pay off in significant improvement of student achievement.

A penny on the dollar would do it. A penny on the dollar would mean $39 million allocated solely to research and development of effective Chapter 1 programs. This is such a small portion of total appropriations for Chapter 1 that it could be raised without reducing services to children. For example, by holding Chapter 1 schools to spring achievement testing only (rather than fall *and* spring), we could save more than enough money.

Can we afford to spend 1% of Chapter 1 funds to make the other 99% pay off? A better question is, Can we afford *not* to do so? No one who looks dispassionately at the results of well-designed evaluations of existing Chapter 1 programs can possibly conclude that we are using Chapter 1 resources to maximum advantage. We know that disadvantaged and low-achieving students can learn. When they fail, it is the system that has failed them. For $500 per student, Chapter 1 in itself may not be able to "break the cycle of poverty," but for $5 per student we can certainly break the cycle of poor programs for students at risk. It is imperative that we do so.

1. John B. Carroll, "The National Assessments in Reading: Are We Misreading the Findings?," *Phi Delta Kappan*, February 1987, pp. 424-28.

2. Launor F. Carter, "The Sustaining Effects Study of Compensatory and Elementary Education," *Educational Researcher*, vol. 13, 1984, pp. 4-13; and Mary M. Kennedy, Beatrice F. Birman, and

Evaluation	Grades Evaluated	Effect Sizes			
Comparison was conducted between experimental groups and randomly selected controls. Groups were studied longitudinally over four years in four schools in the Los Angeles Unified School District.	Grades 1-6 math Grades 3-6 reading Grades 3-6 language	CTBS	Year 1	Year 2	Year 3
		Math			
		Computation	+.36	+.56	+.72
		Concepts	−.02	+.12	+.09
		Applications	+.03	+.12	+.26
		Reading			
		Vocabulary	+.25	+.17	+.58
		Comp.	+.23	−.01	−.24
		Spelling	+.14	+.05	+.14
		Language			
		Mechanics	+.22	+.27	+.25
		Expression	+.11	+.05	+.23
Students were compared to students receiving Chapter 1 pull-out services. Students were randomly assigned in Lafayette Parish, La.	Grades 3-6	CTBS (total math)			+.19
Students were randomly assigned in Merrimack, Mass., and were compared to students receiving Chapter 1 pull-out services.	Grades 2-9	MAT (reading)			+.40
Students were randomly assigned and were compared to students receiving no additional time.	Grade 1	California Achievement Test (reading)			+.58

Randy Demaline, *The Effectiveness of Chapter 1 Services* (Washington, D.C.: Office of Educational Research and Improvement, U.S. Department of Education, 1986).

3. David G. Savage, "Why Chapter 1 Hasn't Made Much Difference," *Phi Delta Kappan*, April 1987, pp. 581-84.

4. Gene V Glass and Mary Lee Smith, *Pull Out in Compensatory Education* (Washington, D.C.: Department of Health, Education, and Welfare, 1977).

5. Advanced Technology, *Local Operation of Title I, ESEA, 1976-1982: A Resource Book* (McLean, Va.: Advanced Technology, 1983).

6. Francis X. Archambault, "Pullout Versus In-Class Instruction in Compensatory Education," paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., 1987.

7. George H. Kimball, John Crawford, and Frank Raia, *Evaluation of the 1984-85 Oklahoma City Public Schools Chapter 1 Program* (Oklahoma City: Oklahoma City Public Schools, 1985).

8. Leonard Granick et al., *Optimizing Chapter 1 Program Effectiveness* (Baltimore: Baltimore City Public Schools, 1984); and Alice G. Pinderhughes et al., *Chapter 1 Services Report FY '86* (Baltimore: Baltimore City Public Schools, 1986).

9. Philip A. Griswold, Kathleen J. Cotton, and Joe B. Hansen, *Effective Compensatory Education Sourcebook* (Washington, D.C.: U.S. Government Printing Office, 1986).

10. For a discussion of methodological problems in the use of fall-to-spring gains, see Roy M. Gabriel et al., *Studying the Sustained Achievement of Chapter 1 Students* (Washington, D.C.: U.S. Department of Education, 1985).

11. Wesley C. Becker and Douglas Carnine, "Direct Instruction: An Effective Approach for Educational Intervention with the Disadvantaged and Low Performers," in Benjamin J. Lahey and Alan E. Kazdin, eds., *Advances in Child Clinical Psychology* (New York: Plenum, 1980), pp. 429-73.

12. Robert E. Slavin, "Team-Assisted Individualization: Combining Cooperative Learning and Individualized Instruction in Mathematics," in Robert E. Slavin et al., eds., *Learning to Cooperate, Cooperating to Learn* (New York: Plenum, 1985), pp. 177-209.

13. Robert J. Stevens, Nancy A. Madden, Robert E. Slavin, and Anna Marie Farnish, "Cooperative Integrated Reading and Composition: Two Field Experiments," *Reading Research Quarterly*, vol. 22, 1987, in press.

14. National Diffusion Network, *Educational Programs That Work*, 13th ed. (Longmont, Colo.: Sopris West, 1987).

15. Gay Su Pinnell, "Reading Recovery: A Pilot Study and Follow-Up of the Effects of an Early Intervention Program in Reading," paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., 1987.

16. Linda Devin-Sheehan, Robert Feldman, and Vernon Allen, "Research on Children Tutoring Children: A Critical Review," *Review of Educational Research*, vol. 46, 1976, pp. 355-85.

17. Marjorie Ragosta, "Computer-Assisted Instruction and Compensatory Education: A Longitudinal Analysis," *Machine-Mediated Learning*, vol. 1, 1983, pp. 97-127.

18. School Improvement Act, 1987, U.S. House of Representatives, Chapter 1 Reauthorization (H.R. 5, Sec. 1012.b.1).

19. Richard Allington and Peter Johnston, "Coordination, Collaboration, and Consistency: The Redesign of Compensatory and Special Education Interventions," paper presented at the Conference on Effects of Alternative Designs in Compensatory Education, Washington, D.C., June 1986.

20. Gabriel et al., *Studying the Sustained Achievement. . . .*

21. Maria Calderon and David D. Marsh, "Applying Research on Effective Bilingual Instruction in a Multi-District Inservice Teacher Training Program," paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., 1987.

"That's the last time I write an essay about what I did on Halloween."

How Valid Are Principals' Judgments of Teacher Effectiveness?

Not very valid, respond the authors. Consequently, we must find some better way of assessing teacher effectiveness – and soon.

BY DONALD M. MEDLEY AND HOMER COKER

NEARLY all decisions about teachers' roles stem from judgments of their effectiveness – judgments usually made by principals. Although the question of whether or not these judgments are valid is a natural and important one, it is rarely asked. The validity of principals' judgments of the effectiveness of the teachers they supervise is generally taken for granted.

However, the small number of studies of the validity of principals' judgments (or of ratings based on them) that have been reported in the literature have yielded consistently negative findings. Each such study has concluded that there is no appreciable agreement between principals' judgments of teachers' effectiveness and the amount students learn.[1]

The implications of this conclusion are grave indeed. How can any effort to in-

DONALD M. MEDLEY (University of Virginia Chapter) is professor emeritus at the University of Virginia, Charlottesville. HOMER COKER (West Georgia College Chapter) is director of the Teacher Evaluation Project and a professor of education at Georgia State University, Atlanta.

Illustration by Kevin Pope

crease the effectiveness of schools in promoting learning succeed if questions concerning which teachers are certified, hired, given tenure, and recognized as outstanding depend, not on how effective they are, but on some unknown factor?

Contemplating such a bleak prospect led us to reconsider the question and to design a new study of the validity of principals' judgments.[2] In doing so, we reviewed the studies cited above and discovered a number of common weaknesses in design and instrumentation that might have prevented them from detecting a relationship between principals' judgments and teachers' effectiveness.

One such weakness arises from the necessity of using a sample of teachers drawn from different schools in order to secure enough teachers to allow a relationship to be detected. This means that judgments made by different principals in different schools must be treated as interchangeable, which they almost certainly are not. It also means that the same achievement test must be used to measure student learning in all the schools and that differences in the objectives sought in different schools must be ignored.

Another weakness we discovered in the studies we examined was that the statistical procedures used to estimate the effectiveness of teachers violated important assumptions. Since these studies were reported, it has been shown that, because of a regression artifact, the procedure most commonly used yields distorted estimates of teacher effectiveness.[3] Methodological advances made it possible to design and conduct a study that would have none of these weaknesses – a study that would yield correlations that would not underestimate the relationship between principals' judgments and teachers' effectiveness, as we suspected the earlier studies had done. We decided to conduct such a study.

We selected a sample of 46 elementary school principals who were employed in systems whose testing programs provided suitable data and who were willing to participate in the study. Principals were asked to judge the effectiveness of those teachers under their supervision whom they felt qualified to judge. The principals recorded their judgments of each teacher's performance in three roles: 1) providing learning experiences that result in the acquisition of fundamental knowledge by pupils; 2) providing learning experiences that lead to good citizenship, personal satisfaction, and self-understanding; and 3) acting as a professional colleague of other teachers, supervisors, and administrators.

Our sample contained 87 groups of elementary school teachers, each made up of teachers of the same grade and subject in the same school whose performance had been judged by the same principal. The total number of teachers for whom completed data were available was 322; the total number of principals was 46.

Pretest and posttest scores of students in each teacher's class were obtained on equivalent forms of the reading test and the arithmetic test administered as part of the regular school testing program. A regression equation was derived for each teacher's class in order to predict posttest scores from pretest scores. These equations were then used to obtain an estimate of the posttest score an average student would be expected to achieve in each teacher's class, and this expected score was used as a measure of the effectiveness of the teacher in that class.

Coefficients of correlation were calculated separately for each of the 87 groups of teachers judged by the same principals, so that 87 separate estimates of the validity of principals' judgments were obtained for each of the two subject-matter areas. Table 1 shows the mean correlations across groups and subjects, by grades and by roles judged.

The correlation coefficient in Table 1 that corresponds most closely to those reported in earlier studies is the overall correlation of .20 with performance of Role 1, a value that is in startlingly close agreement with the earlier findings and can only be interpreted as a confirmation of them.

Table 1 also shows that principals' judgments of teachers' performance of different roles are essentially interchangeable. This finding is a further bit of evidence that the principals' judgments have little to do with teachers' effectiveness in promoting student achievement. It seems to us that the time has come to accept this unpleasant fact and to stop pretending that expert opinions about teacher effectiveness are valid.

The time has come to stop pretending that expert opinions about teacher effectiveness are valid.

If we accept what seems to be the inescapable conclusion that principals' judgments of teacher effectiveness have little or nothing to do with achievement gains, then the first question we must ask is, What do principals' ratings measure? Since these ratings tend to be reliable, they must be measuring some stable teacher characteristic. Whatever that characteristic is, it is of critical importance, since it is the actual basis on which teachers in our schools are recruited, selected, promoted, and recognized as meritorious. It is our guess that these ratings reflect mainly the kind of learning environment a teacher creates and the kind of rapport a teacher has with stu-

TABLE 1.
Mean Correlations Between Principals' Judgments of Teacher Effectiveness in Performing Three Teaching Roles and Achievement Gains of Students

Grade	N	Role 1	Role 2	Role 3	Mean
2	30	.20	.13	.03	.12
3	10	.26	.24	.22	.24
4	12	.16	.10	.05	.10
5	16	.23	.22	.25	.23
6	19	.17	.24	.13	.18
Overall	87	.20	.19	.13	.17

dents – both qualities that are universally regarded as desirable. Why principals' ratings do not correlate with achievement gains is an important question for future research.

If the ability of a teacher to provide "learning experiences that result in pupils' acquisition of fundamental knowledge" is regarded as an important criterion on which to evaluate teachers, then we must answer another question: How can we obtain valid assessments of teacher effectiveness for use in day-to-day decisions?

One might ask whether anything can be done to make principals' ratings more responsive to teacher effectiveness. Can principals be trained to be better judges of teacher effectiveness? The strength of the overall impression that a principal (or anyone else, for that matter) forms of the effectiveness of any teacher is so strong that we doubt that any amount of training can overcome it. But the effort is probably worth making.

WE MUST ALSO ask ourselves whether some alternative measure of teacher effectiveness could be used instead of principals' ratings. The alternative most often suggested as a way of assessing teachers is a "teacher competency test." But repeated efforts to establish the validity of such tests as measures of teacher effectiveness have been no more successful than efforts to validate principals' judgments.[4]

A second alternative that is sometimes proposed is the use of achievement gains of students in the teachers' classes as a measure of teacher effectiveness. In addition to being cumbersome and expensive to implement, teacher evaluations based on student gains have been found to be unreliable.[5] They are, in fact, likely to do more harm than good when used for routine teacher evaluation.[6]

Can some alternative measure of effectiveness be used instead of principals' ratings?

A third alternative is measurement-based teacher evaluation, which would base teacher evaluation on records of classroom performance made by observers trained to record behavior without evaluating it. This alternative is the most promising, but it needs – and should be given – further development.[7]

Meanwhile, state legislatures continue to mandate statewide programs to evaluate teachers. A national committee is now beginning to plan a nationwide system of teacher evaluation. These efforts continue in apparent ignorance of the unsolved problems of teacher evaluation. How much longer can we afford to ignore them?

"I am a 'substitute' teacher, not a 'second-string' teacher."

1. Arvil S. Barr et al., "The Validity of Certain Instruments Employed in the Measurement of Teaching Ability," in Helen M. Walker, ed., *The Measurement of Teaching Efficiency* (New York: Macmillan, 1935), Ch. 4; Wilbur B. Brookover, "The Relation of Social Factors to Teaching Ability," *Journal of Experimental Education*, vol. 13, 1945, pp. 157-65; A. G. Hellfritsch, "A Factor Analysis of Teacher Abilities," *Journal of Experimental Education*, vol. 14, 1945, pp. 101-34; Ronald D. Jones, "The Prediction of Teaching Efficiency from Objective Measures," *Journal of Experimental Education*, vol. 15, 1946, pp. 85-99; C. V. LaDuke, "The Measurement of Teaching Ability," *Journal of Experimental Education*, vol. 14, 1945, pp. 75-100; Leo J. Lins, "The Prediction of Teaching Efficiency," *Journal of Experimental Education*, vol. 15, 1946, pp. 2-60; H. M. Anderson, "A Study of Certain Criteria of Teaching Effectiveness," *Journal of Experimental Education*, vol. 23, 1954, pp. 41-71; and Donald M. Medley and Harold E. Mitzel, "Some Behavioral Correlates of Teacher Effectiveness," *Journal of Educational Psychology*, vol. 50, 1959, pp. 239-46.

2. Donald M. Medley and Homer Coker, "The Accuracy of Principals' Judgments of Teacher Performance," *Journal of Educational Research*, March/April 1987, pp. 242-47.

3. Donald T. Campbell and Albert Erlebacher, "How Regression Artifacts in Quasi-Experimental Evaluations Can Mistakenly Make Compensatory Education Look Harmful," in J. Hellmuth, ed., *The Disadvantaged Child*, Vol. 3 (New York: Breimer/Mazel, 1971).

4. Arvil S. Barr, "Measurement and Prediction of Teacher Efficiency," *Journal of Experimental Education*, vol. 16, 1948, pp. 203-83; and Thomas J. Quirk, Barbara J. Witten, and Susan F. Weinberg, "Review of Studies of the Concurrent and Predictive Validity of the National Teacher Examination," *Review of Educational Research*, vol. 43, 1973, pp. 89-113.

5. Jere Brophy, "Stability in Teacher Effectiveness," *American Educational Research Journal*, vol. 10, 1970, pp. 245-52; and Barak Rosenshine, "The Stability of Teacher Effects upon Student Achievement," *Review of Educational Research*, vol. 40, 1970, pp. 647-62.

6. Robert S. Soar, Homer Coker, and Donald M. Medley, "Teacher Evaluation: A Critique of Currently Used Methods," *Phi Delta Kappan*, December 1983, pp. 239-46.

7. Donald M. Medley, Homer Coker, and Robert S. Soar, *Measurement-Based Evaluation of Teacher Performance* (New York: Longman, 1984). K